

AutoBayesian: Developing Bayesian Networks Based on Text Mining

Sandeep Raghuram¹, Yuni Xia¹, Jiaqi Ge¹, Mathew Palakal¹, Josette Jones¹,
Dave Pecenka², Eric Tinsley², Jean Bandos², and Jerry Geesaman²
1: Indiana University - Purdue University Indianapolis, USA
2: My Health Care Manager, Inc.

Abstract. Bayesian network is a widely used tool for data analysis, modeling and decision support in various domains. There is a growing need for techniques and tools which can automatically construct Bayesian networks from massive text or literature data. In practice, Bayesian networks also need be updated when new data is observed, and literature mining is a very important source of new data after the initial network is constructed. Information closely related to Bayesian network usually includes the causal associations, statistics information and experimental results. However, these associations and numerical results cannot be directly integrated with the Bayesian network. The source of the literature and the perceived quality of research needs to be factored into the process of integration. In this demo, we will present a general methodology and toolkit called AutoBayesian that we developed to automatically build and update a Bayesian network based on the casual relationships derived from text mining.

1 Introduction

A Bayesian network (BN) is a directed acyclic graph whose arcs denote a direct causal influence between parent nodes (causes) and children nodes (effects). A BN is often used in conjunction with statistical techniques as a powerful data analysis and modeling tool. While it can handle incomplete data and uncertainty in a domain, it can also combine prior knowledge with new data or evidence [1].

There are two approaches to construct a BN: knowledge-driven and data-driven. The knowledge-driven approach involves using an expert's domain knowledge to derive the causal associations; and the data driven approach derives the mappings from data which can then be validated by the expert [2]. Data-driven approach has gained much popularity in recent years due to its automated nature and its potential to bring new insights to human being.

2 Demonstration

In this demo, we will show AutoBayesian, a data driven tool developed to build Bayesian network based on the casual relationships derived from text mining [3]. It was developed using Microsoft SQL Server 2009 Express edition and a Bayesian network development tool called NETICA. AutoBayesian system has been tested in geriatrics health care. Figure 1 shows a sample Bayesian network derived

based on the text mining data. This BN is for fall risk evaluation and management for senior patients. In the demo, we will show step by step how AutoBayesian builds a Bayesian network from text mining information and how it interactively updates the BN when new evidences are observed.

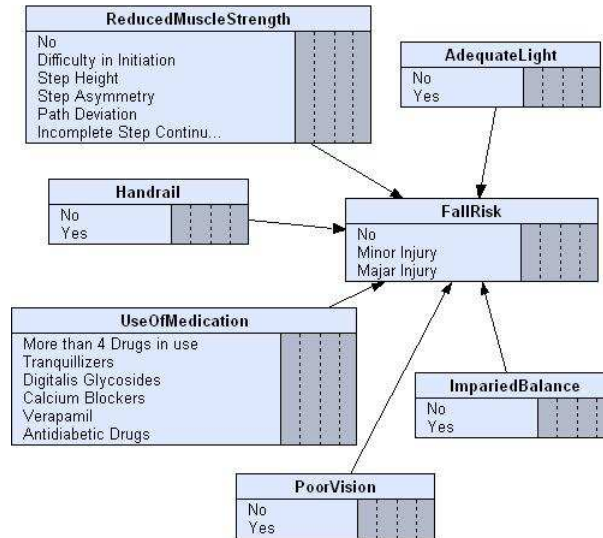


Figure 1. A Sample Bayesian Network Derived from Text

2.1 Derive Confidence Measure

By using existing text mining techniques, causal associations can be extracted from geriatrics health care literature. After the probabilities have been extracted and assessed, we will determine how much confidence we have in the causal associations mined from text. The confidence measure is a score we associate with every causal mapping in the BN based on the confidence we have in asserting that relationship. It quantifies the confidence placed in the causal relationship uncovered by automated methods. In this respect, the two most important parameters we consider are the journal's influence measure and the evidence level of the causal relationship itself. The confidence measure is then computed as a weighted average of the journal's influence measure and the evidence level of the evidence[4].

2.2 Integrate the Causal Mapping with Bayesian Network

Mapping the mined noun phrases to a node in the existing BN is a semantic classification problem and can be solved using information retrieval and/or classification techniques. Using k-nearest neighbor (k-nn) technique, the new noun phrase can be searched in a space containing all the node names. Another method involves use of vector representation of the names of the nodes in the BN. The new noun phrases are also converted into a vector and compared to all the existing vectors to find a match. For a domain which has a large training data, machine learning techniques such as Weight-normalized Complement Naive Bayes (WCNB) will be

used. The process of mapping noun phrases to nodes in a BN has to be highly interactive. Therefore, we also provide an interface so that expert can choose to build the mapping between noun phrases to nodes in a BN, as shown in Figure 2. The system will show two lists, one contains the unmapped keywords and the other contains the available nodes in the BN. The expert can manually build a mapping by choosing a keyword and the corresponding mapping node in the BN and then submit it.

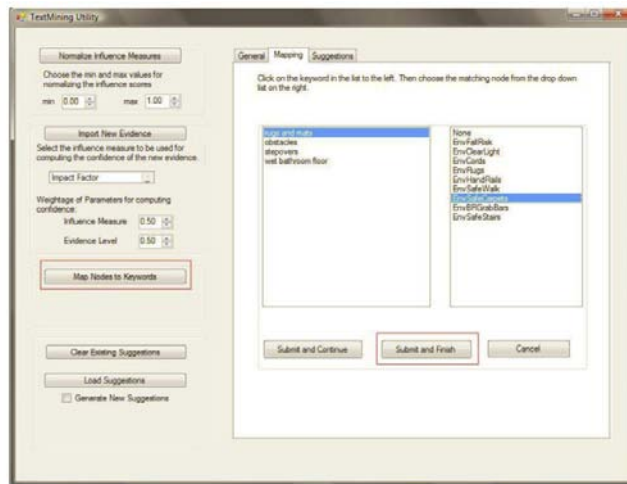


Figure 2. Mapping keywords to Nodes in Bayesian Network

Keyword	Source_Node	Target_Node	Probability	Confidence
	Env Safe Stairways	Environmental Fall Risk	0.74824	0.71790
		Environmental Fall Risk	0.65000	0.74310
./	Env Electrical Cords Clear?	Environmental Fall Risk	0.44620	0.80360
	Env Safe Walkways	Environmental Fall Risk	0.33659	0.77583
	Env Clear and Adequate ...	Environmental Fall Risk	0.37848	0.82555
		Environmental Fall Risk	0.43000	0.74310
		Environmental Fall Risk	0.32000	0.74310
		Environmental Fall Risk	1.00000	0.74310
		Environmental Fall Risk	0.75000	0.63440
		Environmental Fall Risk	0.80000	0.88300
		Environmental Fall Risk	0.75000	0.93300
		History of Arthritis	0.80000	0.65000
*				

Buttons: Review Selected Suggestions, Approve Selected Suggestions

Figure 3. Evidence to be reviewed

Our system consolidates all the evidences and writes out the result into a database table. It identifies all the unique triplets based on the nodes mapped to them and computes the probability and confidence as discussed earlier. The nodes representing cause-effect relation are also written out with the result. If the evidence is new and

has no associated representation in the Bayesian Networks, then the triplet along with its probability and confidence is written out as it is but the fields representing the cause-effect nodes are left null to indicate that it is a new causal association. The generated suggestions are then displayed on the screen for review by the expert. For causal associations already existing in the BN, the previous probability and confidence is displayed to facilitate comparison with the newer values. For causal associations which induce loops in the BN, a message is displayed indicating the same. Once the suggestions are generated and displayed on the screen, the expert can choose to automatically accept the suggestions, or to review them by selecting the interesting suggestion, as shown in Figure 3. The system then displays this suggestion as part of the appropriate Bayesian Network. When both the nodes and the link between them exist, only the conditional probability table needs to be updated.

Figure 4 demonstrates a case when both nodes exist in the network but are not linked causally. The system will create the link if it does not induce any loops in the network. Figure 4 shows the BN after applying the new evidence. As shown in the Figure, the new evidence linking the client's gender and arthritis is applied to the BN.

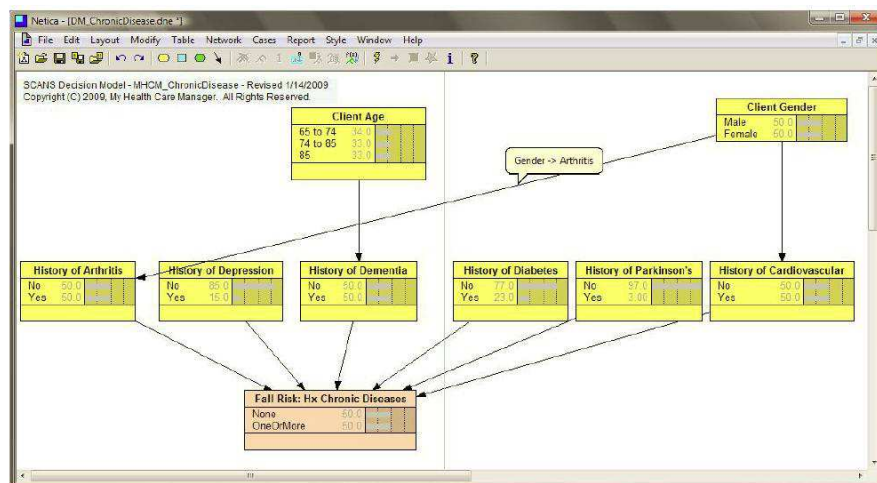


Figure 4. Bayesian Network after Adding a New Link

References

1. S. Nadkarni and P. Shenoy, "A causal mapping approach to constructing bayesian networks," Decision Support Systems, vol. 38, pp. 259–281, 2004.
2. D. Heckerman, "Bayesian networks for data mining," Data Mining and Knowledge Discovery, 1996.
3. S. Raghuram, Y. Xia, M. Palakal, J. Jones, D. Pecenka, E. Tinsley, J. Ban- dos, and J. Geesaman, "Bridging text mining and bayesian networks," Proc. of the Workshop on Intelligent Biomedical Information Systems, 2009.
4. E. Capezuti, D. Zwicker, M. Mezey, and T. Fulmer, Evidence-based geriatric nursing protocols for best practice. Springer, 2008